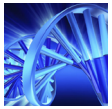
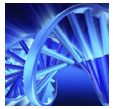


Multiple Sequence Alignment



Sequences

> Yeast YOR020c
mstllksaksivplmdrvlvqrikaqaktasglylpe
knveklnqaevvavgpgftdangnkvvpqvkvgdqvl
ipqfggstiklgnddevilfrdaeilakiakd

> Neurospora crassa
mattvrsvksliplldrvlvqrvkaeaktasgiflpe
ssvkdlneakvlavgpgaldkdgkrlpmgvnagdrvl
ipqyggspvkvgeeeytlfrdseilakiae

> Aspergillus nidulans
msllrnvknlaplldrvlvqrvkpeaktasgiflpes
svkeqneakvlavgpgavdrngqripmgvaagdrvlv
pqfggsplkigeeeyhlfrdseilakine

> Schizosaccharomyces pombe (fission yeast)
matklksaksivplldrilvqrikadtktasgiflpe
ksveklsegrvisvgkggynkegklaqpsvavgdrvl
lpayggsnikvgeeeyslyrdhellaiike

> Mortierella alpina
masritkfsktivpmmdrvlvqrikpqqktasgiyip
ekaqealnegyvvavgkglttqegkvvpselaegdkv
llppyggsvvkvdneelilfreseilakiq

> Crypthecodinium cohnii
matgiakrftplldrvlvqrlkpeaktasglflpesa
akapnyatvlavgpggrtrdgdilpmnvkvgdkvvvp
eyggmtlkfedeefqvfrdadimgilne

> Drosophila melanogaster
maaaikkiipmldriliqraealtktkggivlpekav
gkvlegtvlavgpgtrnastgnhipigvkegdrvllp
efggtkvnlegdqkelflfresdilakle

> Homo sapiens
agqafrkflplfdrvlversaaetvtkggimlpeksq
gkvlqatvvavgsgskgkggeiqpvsvkvgdkvllpe
yggtkvvlddkdyflfrdgxilgky

> Geobacillus stearothermophilus
vlkplgdrvvievieteektasgivlpdtakekpqeg
rvvavgkgrvldsgervapevevgdriifskyagtev
kydgkeylilresdilavig

> Mycobacterium tuberculosis
makvnikpledkilvqaneaetttasglvipdtakek
pqegtvvavgpgrwdedgekripldvaegdtviysky
ggteikyngeeylilsardvlavvsk

> Mus musculus (house mouse)
magqafrkfllfdrvlversaaetvtkggimlpeks
qgkvlqatvvavgsggkgksgeiepvsvkvgdkvllp
eyggtkvvlddkdyflfrdsdilgkyvn

Multiple Sequence Alignment (MSA)

```
Homo          ----AGQAFRKFLPLFDRVLVERSAAETVTKGGIMLPEKSQGKVLQATVVAVGSGSK-GK  55
Mus           ---MAGQAFRKFLLLFDRVLVERSAAETVTKGGIMLPEKSQGKVLQATVVAVGSGGK-GK  56
Drosophila    ----MAAAIKKIIPMLDRILIQRAEALTKTKGGIVLPEKAVGKVLEGTVLAVGPGTRNAS  56
Neurospora    -MATTVRSVKSLIPLLDRVLVQRVKAEAKTASGIFLPESSVKDLNEAKVLAVGPGAL-DK  58
Aspergillus   --MSLLRNVKNLAPLLDRVLVQRVKPEAKTASGIFLPESSVKEQNEAKVLAVGPGAV-DR  57
Crypthecodinium ---MATGIAKRFTPLLDRVLVQRLKPEAKTASGLFLPESAAKAPNYATVLAVGPGGR-TR  56
Yeast         MSTLL-KSAKSIVPLMDRVLVQRIKAQAKTASGLYLPEKNVEKLNQAEVVAVGPGFT-DA  58
Schizosaccharomyces MATKL-KSAKSIVPLLDRILVQRIKADTKTASGIFLPEKSVEKLSEGRVISVGKGGY-NK  58
Mortierella   MASRITKFSKTIVPMMDRVLVQRIKPQQKTASGIYIPEKAQEALNEGYVVAVGKGLT-TQ  59
Geobacillus   ----------VLKPLGDRVVIEVIETEEKTASGIVLPDTAKEKPQEGRVVAVGKGRVLDS  50
Mycobacterium ------MAKVNIKPLEDKILVQANEAETTTASGLVIPDTAKEKPQEGTVVAVGPGRWDED  54
                         :  : *:::::   .   * .*: :*:.       . *::** *

Homo          GGEIQPVSVKVGDKVLLPEYGGTKVVLD--DKDYFLFRDGXILGKY---  99
Mus           SGEIEPVSVKVGDKVLLPEYGGTKVVLD--DKDYFLFRDSDILGKYVN-  102
Drosophila    TGNHIPIGVKEGDRVLLPEFGGTKVNLEGDQKELFLFRESDILAKLE--  103
Neurospora    DGKRLPMGVNAGDRVLIPQYGGSPVKVG--EEEYTLFRDSEILAKIAE-  104
Aspergillus   NGQRIPMGVAAGDRVLVPQFGGSPLKIG--EEEYHLFRDSEILAKINE-  103
Crypthecodinium DGDILPMNVKVGDKVVVPEYGGMTLKFE--DEEFQVFRDADIMGILNE-  102
Yeast         NGNKVVPQVKVGDQVLIPQFGGSTIKLGN-DDEVILFRDAEILAKIAKD  106
Schizosaccharomyces EGKLAQPSVAVGDRVLLPAYGGSNIKVG--EEEYSLYRDHELLAIIKE-  104
Mortierella   EGKVVPSELAEGDKVLLPPYGGSVVKVD--NEELILFRESEILAKIQ--  104
Geobacillus   GE-RVAPEVEVGDRIIFSKYAGTEVKYD--GKEYLILRESDILAVIG--  94
Mycobacterium GEKRIPLDVAEGDTVIYSKYGGTEIKYN--GEEYLILSARDVLAVVSK-  100
                     :  ** :: . :.*  :      .:  :    ::.
```


Why MSA?

- Proteins are often related to a larger group (i.e., a family) of proteins
- Multiple sequence alignment is more sensitive than pairwise alignment for detecting homologs
- MSAs can elucidate conserved residues, motifs, or other functional regions in a protein
- MSA is critical for phylogenetic analysis
  - Selection of sequences
  - Multiple sequence alignment of sequences
  - Tree building
  - Tree evaluation

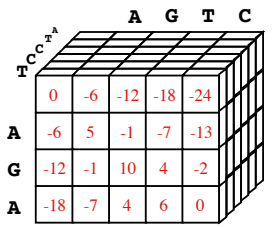
Pairwise Alignment

		A	C	G	T	G	A
	0	-6	-12	-18	-24	-30	-36
A	-6	5	-1	-7	-13	-19	-25
G	-12	-1	1	4	-2	-8	-14
C	-18	-7	4	-2	0	-6	-12
G	-24	-13	-2	9	3	5	-1
T	-30	-19	-8	3	14	8	2
T	-36	-25	-14	-3	8	10	4
A	-42	-31	-20	-9	2	4	15

```
AGCGTTA
|  |||  |
A-CGTGA
```


3-Sequence Alignment

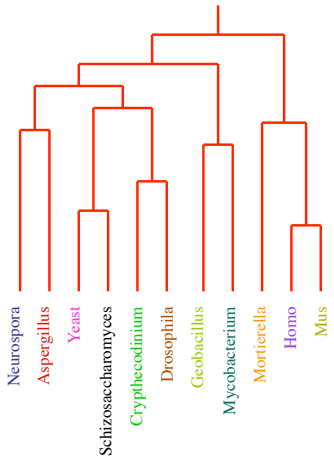
```
AGA
AGTC
TCCTA
```



Pairwise Alignment Scores

	Yeast	Neurospora	Aspergillus	Schizosacchrmyces	Mortierella	Crypthecodinium	Drosophila	Homo	Geobacillus	Mycobacterium	Mus
Yeast	49	46	78	45	55	54	44	38	37	42	
Neurospora		52	41	40	43	46	44	41	39	43	
Aspergillus			43	48	45	45	40	40	38	39	
Schizosaccharomyces				42	53	55	41	41	40	40	
Mortierella					43	46	40	43	38	39	
Crypthecodinium						61	43	34	36	45	
Drosophila							49	42	36	49	
Homo								37	32	93	
Geobacillus									59	38	
Mycobacterium										32	
Mus											

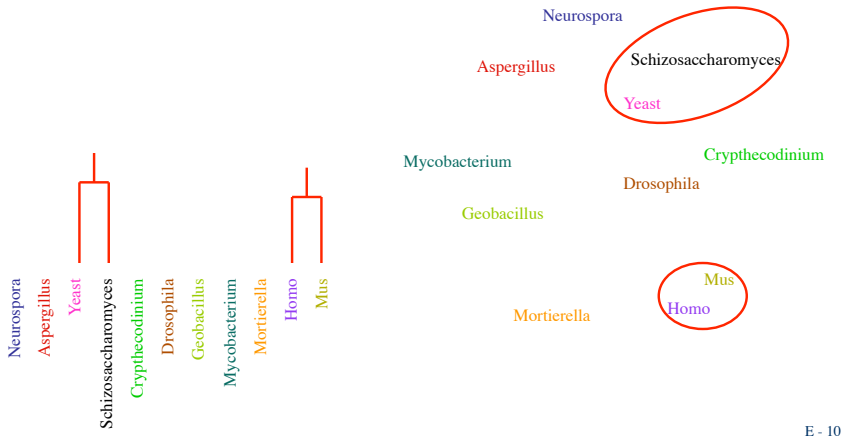
Construct a Guide Tree



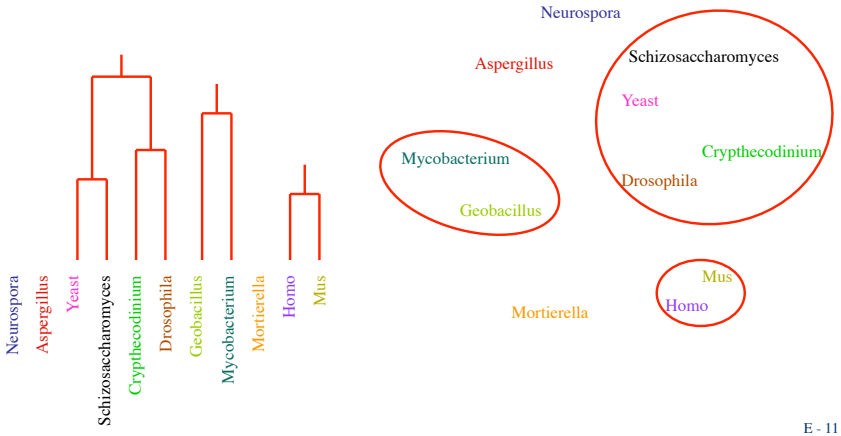
Unweighted Pair Group Method with Arithmetic mean (UPGMA)

- Assume each organism is its own group
- Repeat the following step
  - Merge together the two closest groups

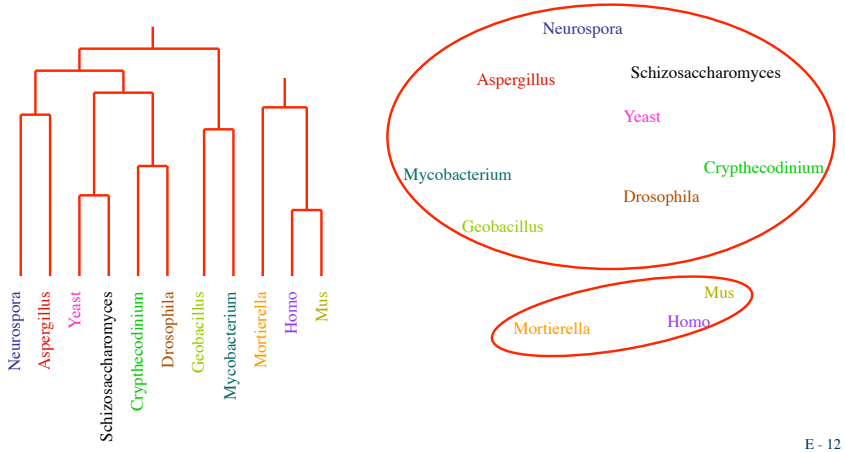
Unweighted Pair Group Method with Arithmetic mean (UPGMA)



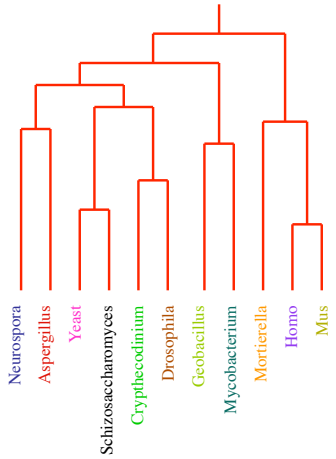
Unweighted Pair Group Method with Arithmetic mean (UPGMA)



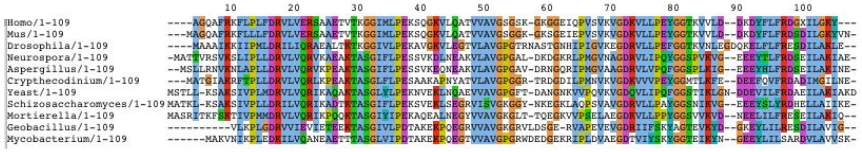
Unweighted Pair Group Method with Arithmetic mean (UPGMA)



Guide Tree



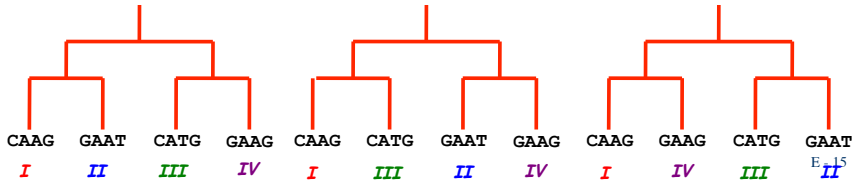
Multiple Sequence Alignment



Constructing a Phylogenetic Tree

When analyzing a set of data, there are *many* possible phylogenies to consider. We would like to identify a good (the best) phylogeny.

I	**CAAG**
II	**GAAT**
III	**CATG**
IV	**GAAG**



Constructing a Phylogenetic Tree

Character based methods are one type of approach for constructing a phylogenetic tree.

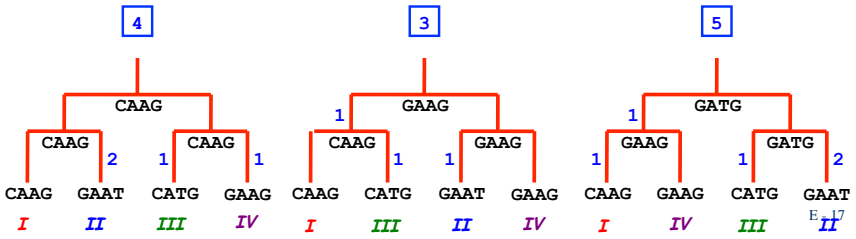
Character based methods are often based on the idea of *maximum parsimony*.

" IT IS VAIN TO DO WITH MORE WHAT CAN BE DONE WITH FEWER"
OR
Principle of parsimony
OR
...smallest number of evolutionary changes...

Optimality criterion: The 'most parsimonious' tree is the one that requires the fewest number of evolutionary events (*e.g.*, nucleotide substitutions, amino acid replacements) to explain the sequences.

Character Based Methods

I	CAAG
II	GAAT
III	CATG
IV	GAAG



Maximum Parsimony Methodology

Step 1: Identify informative sites

Sites with at least two different characters at the site, each of which is represented in at least two of the sequences

	Site								
Seq.	1	2	3	4	5	6	7	8	9
I	A	A	G	A	G	T	T	C	A
II	A	G	C	C	G	T	T	C	T
III	A	G	A	T	A	T	C	C	A
IV	A	G	A	G	A	T	C	C	T

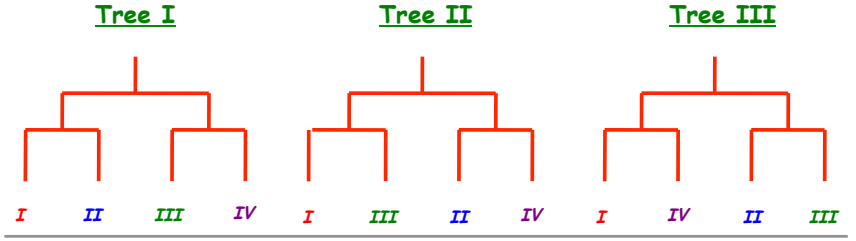
Maximum Parsimony Methodology

Step 1: Identify informative sites

Sites with at least two different characters at the site, each of which is represented in at least two of the sequences

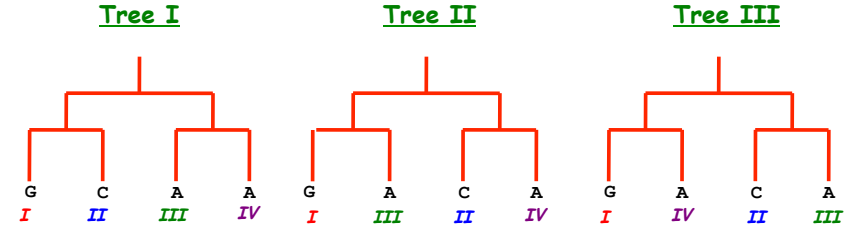
	Site								
Seq.	1	2	3	4	5	6	7	8	9
I	A	A	G	A	G	T	T	C	A
II	A	G	C	C	G	T	T	C	T
III	A	G	A	T	A	T	C	C	A
IV	A	G	A	G	A	T	C	C	T

Sites Where All Trees Require the Same Number of Changes Are Not Informative



	Site								
Seq.	1	2	3	4	5	6	7	8	9
I	A	A	G	A	G	T	T	C	A
II	A	G	C	C	G	T	T	C	T
III	A	G	A	T	A	T	C	C	A
IV	A	G	A	G	A	T	C	C	T

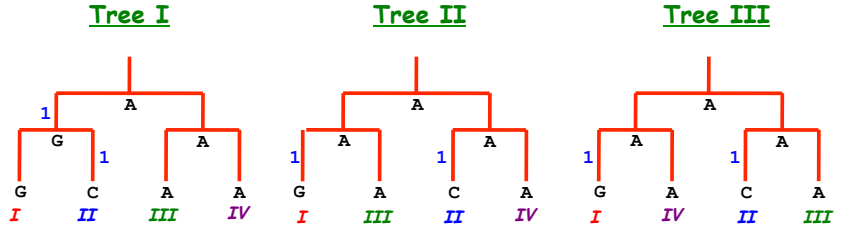
Sites Where All Trees Require the Same Number of Changes Are Not Informative



Site

Seq.	1	2	3	4	5	6	7	8	9
I	A	A	G	A	G	T	T	C	A
II	A	G	C	C	G	T	T	C	T
III	A	G	A	T	A	T	C	C	A
IV	A	G	A	G	A	T	C	C	T

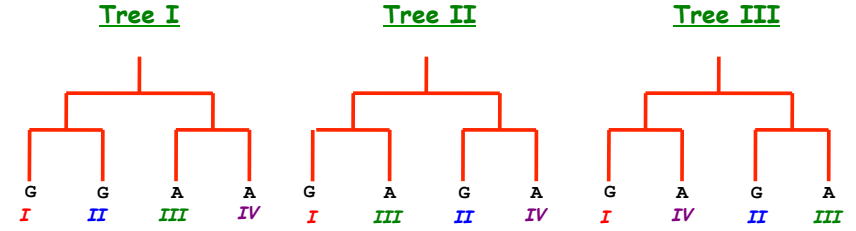
Sites Where All Trees Require the Same Number of Changes Are Not Informative



Site

Seq.	1	2	3	4	5	6	7	8	9
I	A	A	G	A	G	T	T	C	A
II	A	G	C	C	G	T	T	C	T
III	A	G	A	T	A	T	C	C	A
IV	A	G	A	G	A	T	C	C	T

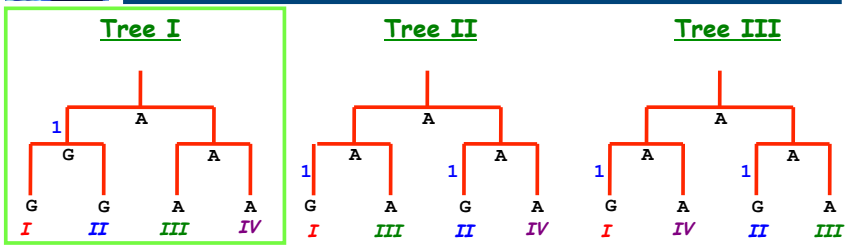
Maximum Parsimony Analyzes Sites At Which One Substitution Model Requires Fewer Changes



Site

Seq.	1	2	3	4	5	6	7	8	9
I	A	A	G	A	G	T	T	C	A
II	A	G	C	C	G	T	T	C	T
III	A	G	A	T	A	T	C	C	A
IV	A	G	A	G	A	T	C	C	T

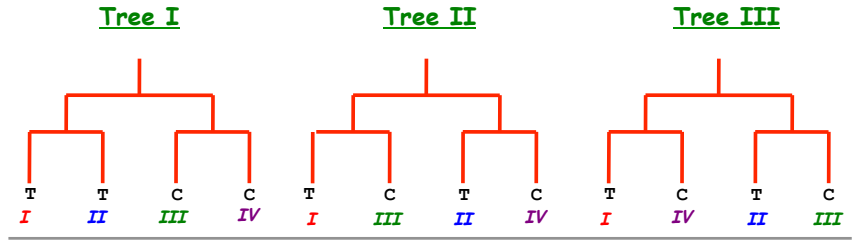
Maximum Parsimony Analyzes Sites At Which One Substitution Model Requires Fewer Changes



Site

Seq.	1	2	3	4	5	6	7	8	9
I	A	A	G	A	G	T	T	C	A
II	A	G	C	C	G	T	T	C	T
III	A	G	A	T	A	T	C	C	A
IV	A	G	A	G	A	T	C	C	T

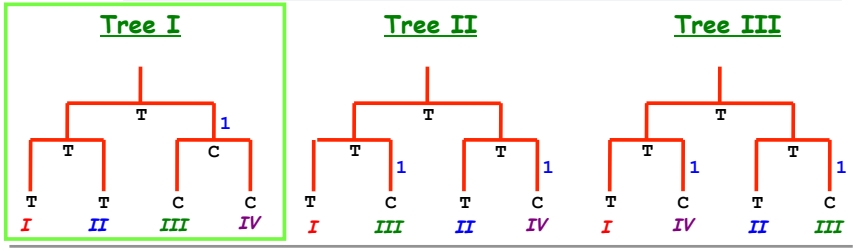
Maximum Parsimony Analyzes Sites At Which One Substitution Model Requires Fewer Changes



Site

Seq.	1	2	3	4	5	6	7	8	9
I	A	A	G	A	G	T	T	C	A
II	A	G	C	C	G	T	T	C	T
III	A	G	A	T	A	T	C	C	A
IV	A	G	A	G	A	T	C	C	T

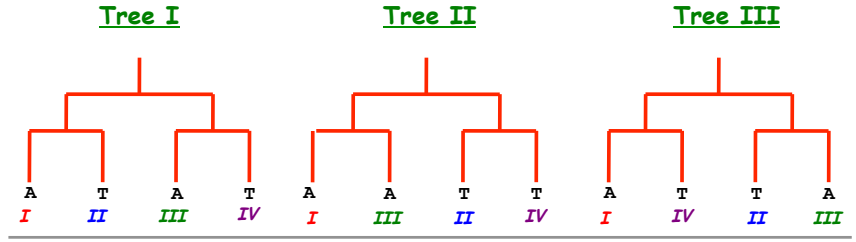
Maximum Parsimony Analyzes Sites At Which One Substitution Model Requires Fewer Changes



Site

Seq.	1	2	3	4	5	6	7	8	9
I	A	A	G	A	G	T	T	C	A
II	A	G	C	C	G	T	T	C	T
III	A	G	A	T	A	T	C	C	A
IV	A	G	A	G	A	T	C	C	T

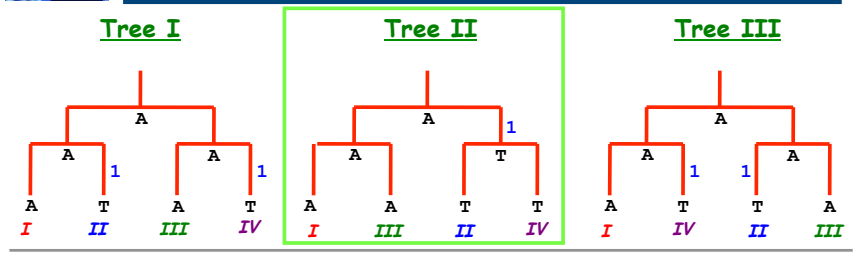
Maximum Parsimony Analyzes Sites At Which One Substitution Model Requires Fewer Changes



Site

Seq.	1	2	3	4	5	6	7	8	9
I	A	A	G	A	G	T	T	C	A
II	A	G	C	C	G	T	T	C	T
III	A	G	A	T	A	T	C	C	A
IV	A	G	A	G	A	T	C	C	T

Maximum Parsimony Analyzes Sites At Which One Substitution Model Requires Fewer Changes



Site

Seq.	1	2	3	4	5	6	7	8	9
I	A	A	G	A	G	T	T	C	A
II	A	G	C	C	G	T	T	C	T
III	A	G	A	T	A	T	C	C	A
IV	A	G	A	G	A	T	C	C	T

Maximum Parsimony Methodology

Step 2: Calculate minimum number of substitutions at each informative site

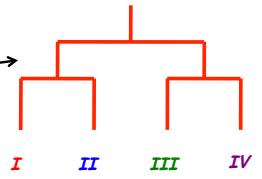
Step 3: Sum number of changes at each informative site for each possible tree

The tree with the least number of total changes is the most parsimonious tree

Number of Changes at Each Informative Site

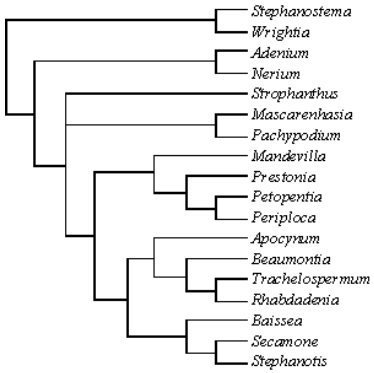
	5	7	9	Σ
Tree I	1	1	2	4
Tree II	2	2	1	5
Tree III	2	2	2	6

Tree I

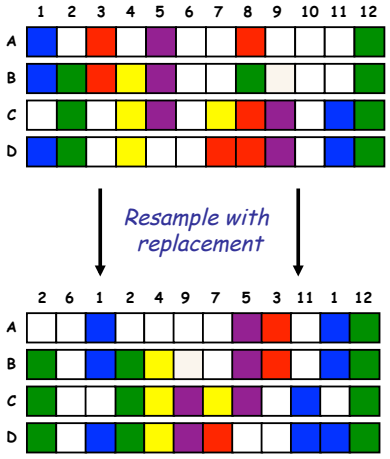


How Confident Are We In Our Tree?

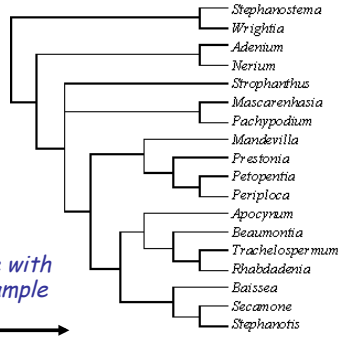
Bootstrapping: Given a particular tree, how consistently does a tree-building algorithm find that branching order using a randomly sampled version of the original dataset?



Random Sampling



Build tree with pseudosample



Bootstrapping

How reliable is our tree?

Repeat the following many times:

- Generate a new set of sequences by randomly sampling the original sequences
- Construct a tree for the new set of randomly sampled sequences
- Compare the new tree (based on the random sequences) with the original tree (based on the original sequences) and determine how many times the same tree structures were recovered

Adenium and *Nerium* were grouped together in 98% of the trees built from randomly sampled sequences